3578–3580 Nucleic Acids Research, 1994, Vol. 22, No. 17

The SWISS-PROT protein sequence data bank: current status

Amos Bairoch* and Brigitte Boeckmann[1]
Department of Medical Biochemistry, University of Geneva, 1 rue Michel Servet, 1211 Geneva 4, Switzerland and [1]European Molecular Biology Laboratory, Heidelberg, Germany

ABSTRACT

SWISS-PROT [1] is an annotated protein sequence database established in 1986 and maintained collaboratively, since 1988, by the Department of Medical Biochemistry of the University of Geneva and the EMBL Data Library [2]. The SWISS-PROT protein sequence data bank consist of sequence entries. Sequence entries are composed of different lines types, each with their own format. For standardization purposes the format of SWISS-PROT [3] follows as closely as possible that of the EMBL Nucleotide Sequence Database. A sample SWISS-PROT entry is shown in Figure 1.

The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria:

a) Annotation

In SWISS-PROT, as in most other sequence databases, two classes of data can be distinguished: the core data and the annotation. For each sequence entry the core data consists of the sequence data; the citation information (bibliographical references) and the taxonomic data (description of the biological source of the protein) while the annotation consists of the description of the following items:

- Function(s) of the protein
- Post-translational modification(s). For example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.
- Domains and sites. For example calcium binding regions, ATP-binding sites, zinc fingers, homeobox, kringle, etc.
- Secondary structure
- Quaternary structure
- Similarities to other proteins
- Disease(s) associated with deficiencie(s) in the protein
- Sequence conflicts, variants, etc.

We try to include as much annotation information as possible in SWISS-PROT. To obtain this information we use, in addition to the publications that report new sequence data, review articles to periodically update the annotations of families or groups of proteins. We also make use of external experts, who have been recruited to send us their comments and updates concerning specific groups of proteins.

We believe that our having systematic recourse both to publications other than those reporting the core data and to subject referees represents a unique and beneficial feature of SWISS-PROT.

In SWISS-PROT, annotation is mainly found in the comment lines (CC), in the feature table (FT) and in the keyword lines (KW). Most comments are classified by 'topics'; this approach permits the easy retrieval of specific categories of data from the database.

b) Minimal redundancy

Many sequence databases contain, for a given protein sequence, separate entries which correspond to different literature reports. In SWISS-PROT we try as much as possible to merge all these data so as to minimize the redundancy of the database. If conflicts exist between various sequencing reports, they are indicated in the feature table of the corresponding entry.

c) Integration with other databases

It is important to provide the users of biomolecular databases with a degree of integration between the three types of sequence-related databases (nucleic acid sequences, protein sequences and protein tertiary structures) as well as with specialized data collections. SWISS-PROT is currently cross-referenced with twelve different databases. Cross-references are provided in the form of pointers to information related to SWISS-PROT entries and found in data collections other than SWISS-PROT. For example the sample sequence shown in Figure 1 contains DR (Data bank Reference) lines that point to EMBL, PIR, PDB, OMIM, and PROSITE. In this particular example it is therefore possible to retrieve the nucleic acid sequence(s) that encodes for that protein (EMBL), the X-ray crystallographic atomic coordinates (PDB), the description of genetic disease(s) associated with that protein (OMIM), or the pattern specific for that family of proteins (PROSITE).

RECENT DEVELOPMENTS

a) Model organisms

We have selected a number of organisms that are the target of genome sequencing and/or mapping projects and for which we intend to:

- Be as complete as possible. All sequences available at a given time should be immediately included in SWISS-PROT. This also includes sequence corrections and updates.
- Provide a high level of annotations.
- Cross-references to specialized database(s) that contain, among other data, some genetic information about the genes that code for these proteins.
- Provide specific indices or documents.

*To whom correspondence should be addressed

The organisms currently selected are:

- *Caenorhabditis elegans* (worm)
- *Dictyostelium discoideum* (slime mold)
- *Drosophila melanogaster* (fly)
- *Escherichia coli*
- *Homo sapiens* (human)
- Saccharomyces cerevisiae (budding yeast)

Organism	Database	Index file cross-referenced	Number of sequences
C.elegans	WormPep	CELEGANS.TXT	679
D.discoideum	DictyDB	DICTY.TXT	1983
D.melanogaster	FlyBase	In preparation	633
E.coli	EcoGene	ECOLI.TXT	2674
H.sapiens	MIM	MIMTOSP.TXT	2862
S.cerevisiae	LISTA (in preparation)	YEAST.TXT	1951

b) Documentation files

SWISS-PROT is distributed with a large number of documentation files. Some of these files have been available for a long time (the user manual, release notes, the various indices for authors, citations, keywords, etc.), but many have been created recently and we are continuously adding new files. The following table list all the documents that are currently available or that will be added in the next few months.

USERMAN .TXT	User manual
RELNOTES.TXT	Release notes
SHORTDES.TXT	Short description of entries in SWISS-PROT
JOURLIST.TXT	List of abbreviations for journals cited
KEYWLIST.TXT	List of keywords in use
SPECLIST.TXT	List of organism identification codes
ACINDEX .TXT	Accession number index
AUTINDEX.TXT	Author index
CITINDEX.TXT	Citation index
KEYINDEX.TXT	Keyword index
SPEINDEX.TXT	Species index
7TMRLIST.TXT	List of 7-transmembrane G-linked receptors entries
CDLIST .TXT	CD nomenclature for surface proteins of human leucocytes
CELEGANS.TXT	Index of Caenorhabditis elegans entries and their corresponding gene designations and WormPep cross-references
DICTY .TXT	Index of Dictyostelium discoideum entries and their corresponding gene designations and DictyDB cross-references
EC2DTOSP.TXT	Index of Escherichia coli Gene-protein database entries referenced in SWISS-PROT
ECOLI .TXT	Index of Escherichia coli K12 chromosomal entries and their corresponding EcoGene cross-reference
EMBLTOSP.TXT	Index of EMBL Database entries referenced in SWISS-PROT
EXPERTS .TXT	List of on-line experts for PROSITE and SWISS-PROT
GLYCOSYL.TXT	Index of glycosyl hydrolases classified by families on the basis of sequence similarities [*]
HOXLIST .TXT	Vertebrate homeobox proteins: nomenclature and index
MIMTOSP .TXT	Index of MIM entries referenced in SWISS-PROT
NOMLIST .TXT	List of nomenclature related references for proteins
PDBTOSP .TXT	Index of Brookhaven PDB entries referenced in SWISS-PROT
PLASTID .TXT	List of chloroplast and cyanelle encoded proteins
RESTRIC .TXT	List of restriction enzymes and methylases entries
RIBOSOMP.TXT	Index of ribosomal proteins classified by families on the basis of sequence similarities [*]
YEAST .TXT	Index of Saccharomyces cerevisiae entries and their corresponding gene designations
YEAST11 .TXT	Yeast Chromosome XI entries

[*] Will be available starting with release 30 in October 1994.

```
ID   ARSA_HUMAN     STANDARD;      PRT;   507 AA.
AC   P15289;
DT   01-APR-1990 (REL. 14, CREATED)
DT   01-FEB-1991 (REL. 17, LAST SEQUENCE UPDATE)
DT   01-JUN-1994 (REL. 29, LAST ANNOTATION UPDATE)
DE   ARYLSULFATASE A PRECURSOR (EC 3.1.6.8) (ASA) (CEREBROSIDE-SULFATASE).
GN   ARSA.
OS   HOMO SAPIENS (HUMAN).
OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC   EUTHERIA; PRIMATES.
RN   [1]
RP   SEQUENCE FROM N.A.
RM   90361046
RA   KREYSING J., VON FIGURA K., GIESELMANN V.;
RL   EUR. J. BIOCHEM. 191:627-631(1990).
RN   [2]
RP   SEQUENCE FROM N.A.
RM   89093115
RA   STEIN C., GIESELMANN V., KREYSING J., SCHMIDT B., POHLMANN R.,
RA   WAHED A., MEYER H.E., O'BRIEN J.S., VON FIGURA K.;
RL   J. BIOL. CHEM. 264:1252-1259(1989).
RN   [3]
RP   SEQUENCE OF 19-33 AND 434-479, AND SUBUNITS.
RM   92338230
RA   FUJII T., KOBAYASHI T., HONKE K., GASA S., ISHIKAWA M., SHIMIZU T.,
RA   MAKITA A.;
RL   BIOCHIM. BIOPHYS. ACTA 1122:93-98(1992).
RN   [4]
RP   VARIANT MLD GLN-84.
RM   92344341
RA   KAPPLER J., VON FIGURA K., GIESELMANN V.;
RL   ANN. NEUROL. 31:256-261(1992).
RN   [5]
RP   VARIANT MLD PHE-96.
RM   91328147
RA   GIESELMANN V., FLUHARTY A.L., TONNESEN T., VON FIGURA K.;
RL   AM. J. HUM. GENET. 49:407-413(1991).
RN   [6]
RP   VARIANT MLD ASP-99.
RM   91206410
RA   KONDO R., WAKAMATSU N., YOSHINO H., FUKUHARA N., MIYATAKE T.,
RA   TSUJI S.;
RL   AM. J. HUM. GENET. 48:971-978(1991).
RN   [7]
RP   VARIANT MLD SER-122.
RM   94063853
RA   HONKE K., KOBAYASHI T., FUJII T., GASA S., XU M., TAKAMARU Y.,
RA   KONDO R., TSUJI S., MAKITA A.;
RL   HUM. GENET. 92:451-456(1993).
RN   [8]
RP   VARIANT MLD ARG-245.
RM   93319632
RA   HASEGAWA Y., KAWAME H., ETO Y.;
RL   DNA CELL BIOL. 12:493-498(1993).
RN   [9]
RP   VARIANT MLD MET-274.
RM   94004907
RA   HARVEY J.S., NELSON P.V., CAREY W.F., ROBERTSON E.F., MORRIS C.P.;
RL   HUM. MUTAT. 2:261-267(1993).
RN   [10]
RP   VARIANT MLD SER-309.
RM   93318834
RA   KREYSING J., BOHNE W., BOSENBERG C., MARCHESINI S., TURPIN J.C.,
RA   BAUMANN M., VON FIGURA K., GIESELMANN V.;
RL   AM. J. HUM. GENET. 53:339-346(1993).
RN   [11]
RP   VARIANT SER-350.
RM   90083282
RA   GIESELMANN V., POLTEN A., KREYSING J., VON FIGURA K.;
RL   PROC. NATL. ACAD. SCI. U.S.A. 86:9436-9440(1989).
RN   [12]
RP   VARIANT MLD LEU-426.
RM   93202658
RA   BARTH M.L., FENSOM A., HARRIS A.;
RL   HUM. GENET. 91:73-77(1993).
CC   -!- FUNCTION: HYDROLYSES CEREBROSIDE SULFATE.
CC   -!- CATALYTIC ACTIVITY: A CEREBROSIDE 3-SULFATE + H(2)O = A
CC       CEREBROSIDE + SULFATE.
CC   -!- SUBCELLULAR LOCATION: LYSOSOMAL.
CC   -!- SUBUNIT: EXISTS BOTH AS A SIGHLE CHAIN OF 58 KD (COMPONENT A)
CC       OR AS A CHAIN OF 50 KD (COMPONENT B) LINKED BY DISULFIDE BOND(S)
CC       TO A 7 KD CHAIN (COMPONENT C).
CC   -!- DISEASE: DEFICIENCIES IN ARSA ARE A CAUSE OF METACHROMATIC
CC       LEUCODYSTROPHY (MLD); A DISEASE CHARACTERIZED BY INTRALYSOSOMAL
CC       STORAGE OF CEREBROSIDE SULFATE. THREE FORMS OF THE DISEASE CAN BE
CC       DISTINGUISHED ACCORDING TO THE AGE AT ONSET: LATE-INFANTILE,
CC       JUVENILE AND ADULT.
CC   -!- SIMILARITY: TO OTHER SULFATASES.
DR   EMBL; X52150; HSARYLA.
DR   EMBL; X52151; HSARYA.
DR   EMBL; J04593; HSASFA.
DR   PIR; A32207; A32207.
DR   PIR; S11031; S11031.
DR   MIM; 250100; 11TH EDITION.
DR   PROSITE; PS00149; SULFATASE_2.
DR   PROSITE; PS00523; SULFATASE_1.
KW   HYDROLASE; SIGNAL; GLYCOPROTEIN; LYSOSOME; DISEASE MUTATION;
KW   METACHROMATIC LEUCODYSTROPHY; SPHINGOLIPID METABOLISM; POLYMORPHISM.
FT   SIGNAL       1     18
FT   CHAIN       19    507       ARYLSULFATASE A.
FT   CHAIN       19    444       COMPONENT B.
FT   CHAIN      448    507       COMPONENT C.
FT   ACT_SITE   125    125       POTENTIAL.
FT   CARBOHYD   158    158       POTENTIAL.
FT   CARBOHYD   184    184       POTENTIAL.
FT   CARBOHYD   350    350
FT   VARIANT     84     84       R -> Q (IN MLD; LATE-ONSET).
FT   VARIANT     96     96       S -> F (IN MLD; LATE-INFANTILE TYPE).
FT   VARIANT     99     99       G -> D (IN MLD; ADULT TYPE).
FT   VARIANT    122    122       G -> S (IN MLD; ADULT TYPE).
FT   VARIANT    245    245       G -> R (IN MLD; LATE-INFANTILE TYPE).
FT   VARIANT    274    274       T -> M (IN MLD; LATE-INFANTILE TYPE).
FT   VARIANT    309    309       G -> S (IN MLD; LATE-INFANTILE TYPE;
FT                               13% OF NORMAL ACTIVITY).
FT   VARIANT    426    426       P -> L (IN MLD).
FT   VARIANT    350    350       N -> S.
SQ   SEQUENCE   507 AA;  53588 MW;  1281483 CN;
     MGAPRSLLLA LAAGLAVARP PNIVLLFADD LGYGDLGCYG HPSSTTPNLD QLAAGGLRFT
     DFYVPVSLCT PSRAALLTGR LPVRMGMYPG VLVPSSRGGL PLEEVTVAEV LAARGYLTGM
     AGKWHLGVGP EGAFLPPHQG FHRFLGIPYS HDQGPCQNLT CFPPATPCDG GCDQGLVPIP
     LLANLSVEAQ PPWLPGLEAR YMAFAHDLMA DAQRQDRPFF LYYASHHTHY PQFSGQSFAE
     RSGRGPFGDS LMELDAAVGT LMTAIGDLGL LEETLVIFTA DNGPETMRMS RGGCSGLLRC
     GKGTTYEGGV REPALAFWPG HIAPGVTHEL ASSLDLLPTL AALAGAPLPN VTLDGFDLSP
     LLLGTGKSPR QSLFFYPSYP DEVRGVFAVR TGKYKAHFFT QGSAHSDTTA DPACHASSSL
     TAHEPPLLYD LSKDPGENYN LLGGVAGATP EVLQALKQLQ LLKAQLDAAV TFGPSQVARG
     EDPALQICCH PGCTPRPACC HCPDPHA
//
```

Figure 1. A sample entry from SWISS-PROT.

c) New cross-references

We have recently added cross-references that link SWISS-PROT to the following databases:

- The G-protein−coupled receptor database (GCRDb) prepared by Lee Frank Kolakowski at the Massachusetts General Hospital Renal Unit [4].
- The Maize Genome Database (MaizeDB) developed by the USDA-ARS Maize Genome Project as part of the National Agricultural Library's Plant Genome Research Program.
- The WormPep database prepared by Richard Durbin and Eric Sonnhammer from the MRC Laboratory of Molecular Biology and Sanger Center at Hinxton Hall, Cambridge.

- The DictyDb database prepared by Douglas W. Smith and Bill Loomis from the University of California, San Diego (UCSD).

d) Human genetic diseases

We have continued to extend the coverage of information concerning human genetic diseases and of their characterization at the molecular level. All sequence entries that contain, in their feature table, information on one or more disease-causing mutations are tagged with the keyword 'DISEASE MUTATION'. In the comment lines (CC) we have generalized the use of the 'DISEASE' topic to describe the disease(s) associated with a given protein.

A new comment line topic, 'POLYMORPHISM', has been introduced to describe information concerning polymorphisms and alleles (see the example below).

```
CC   -!- POLYMORPHISM: THERE ARE TWO ALLELES: C3S (C3 SLOW),
CC   THE MOST COMMON ALLELE IN ALL RACES AND C3F (C3 FAST),
CC   RELATIVELY FREQUENT IN CAUCASOIDS, LESS COMMON IN
CC   BLACK AMERICAIN, EXTREMELY RARE IN ORIENTALS.
```

More than a thousand disease-causing point mutations were added to various SWISS-PROT entries.

PRACTICAL INFORMATION

a) Content of the current release

Release 29.0 of SWISS-PROT (June 1994) contains 38,303 sequence entries, comprising 13,464,014 amino acids abstracted from 36,636 references. The data file (sequences and annotations) requires 70 Mb of disk storage space. The documentation and index files require about 22 Mb of disk space. No restrictions are placed on use or redistribution of the data.

b) How to obtain SWISS-PROT

SWISS-PROT is distributed on magnetic tape and on CD-ROM by the EMBL Data Library. The CD-ROM contains both SWISS-PROT and the EMBL Nucleotide Sequence Database as well as other data collections and some database query and retrieval software for MS-DOS and Apple MacIntosh computers. For all enquiries regarding the subscription and distribution of SWISS-PROT one should contact:

EMBL Data Library
European Molecular Biology Laboratory
Postfach 10.22.09, Meyerhofstrasse 1
D-69012 Heidelberg, Germany
Telephone: (+49 6221) 387 258
Telefax: (+49 6221) 387 519 or 387 306
Electronic network address: datalib@EMBL-heidelberg.de

Individual sequence entries can be obtained from the EMBL File Server [5]. Detailed instructions on how to make the best use of this service, and in particular on how to obtain protein sequences, can be obtained by sending to the network address netserv@EMBL-heidelberg.de the following message:

HELP
HELP PROT

If you have access to a computer system linked to the Internet you can obtain SWISS-PROT using FTP (File Transfer Protocol), from the following file servers:

EMBL anonymous FTP server
Internet address: ftp.EMBL-heidelberg.de (or 192.54.41.33)
NCBI Repository (National Library of Medicine, NIH, Washington D.C., U.S.A.)
Internet address: ncbi.nlm.nih.gov (130.14.20.1)
ExPASy (Expert Protein Analysis System) server, University of Geneva, Switzerland
Internet address: expasy.hcuge.ch (129.195.254.61)
National Institute of Genetics (Japan) FTP server
Internet address: ftp.nig.ac.jp (133.39.16.66)

c) Interactive access to PROSITE

You can browse through SWISS-PROT using various Internet Gopher servers that specialize in biosciences (biogophers) [6]. Gopher is a distributed document delivery service that allows a neophyte user to access various types of data residing on multiple hosts in a seamless fashion.

SWISS-PROT is currently available on the ExPASy World-Wide Web (WWW) molecular biology server [7]. WWW, which originated at CERN in Geneva, is a global information retrieval system merging the power of world-wide networks, hypertext and multimedia. Through hypertext links, it gives access to documents and information (including images, movies and sound) available on thousands of servers around the world, using network protocols such as FTP, WAIS, Gopher, X500, etc. as well as the WWW specific HyperText Transfer Protocol (HTTP). To access this server (or any other WWW server), one needs a WWW browser. Public domain browsers exist for a variety of computer systems, including Unix, MS-Windows and Macintoshes. One popular browser available for all three platforms is Mosaic, developed at the National Center for Supercomputing Applications (NCSA) of the University of Illinois at Champaign. It may be obtained by anonymous ftp from ftp.ncsa.uiuc.edu, in the directories /Mosaic, respectively /PC and /Mac. Using a WWW browser, one has access to all the hypertext documents stored on the ExPASy server (as well as other WWW servers).

The ExPASy WWW server may be accessed through its Uniform Resource Locator (URL - the addressing system defined in WWW), which is:

http://expasy.hcuge.ch/

d) Release frequency

The present distribution frequency is four releases per year. Weekly updates are also available; these updates are available by anonymous FTP. Three files are updated every week:

new_seq.dat Contains all the new entries since the last full release.
upd_seq.dat Contains the entries for which the sequence data has been updated since the last release.
upd_ann.dat Contains the entries for which one or more annotation fields have been updated since the last release.

These files are available on the EMBL, NCBI and Expasy servers, whose Internet addresses are listed above.

REFERENCES

1. Bairoch A., Boeckmann B. Nucleic Acids Res. 20:2019–2022(1992).
2. Rice C.M., Fuchs R., Higgins D.G., Stoehr P.J., Cameron G.N. Nucleic Acids Res. 21:2967–2971(1993).
3. Bairoch A. SWISS-PROT protein sequence data bank user manual, Release 29 of June 1994.
4. Kolakowski L.F. Jr. Receptors Channels In press(1994).
5. Stoehr P.J., Omond R.A. Nucleic Acids Res. 17:6763–6764(1989).
6. Gilbert D. Trends Biochem. Sci. 18:107–108(1993).
7. Appel R.D., Bairoch A., Hochstrasser D.F. Trends Biochem. Sci. 19:258–260(1994).